OFFICE OF PUBLIC INSTRUCTION
PO BOX 202501
HELENA MT 59620-2501
www.opi.state.mt.us
(406) 444-3095
(888) 231-9393
(406) 444-0169 (TTY)

Linda McCulloch
Superintendent

Transcript of Video Broadcast of Criterion-Referenced Test (CRT) Questions and Answers Presentation, November 16, 2004

Moderator: Judy Snow, Assessment Director

Panel Members: Brian Hilton, Carol Wicker, Bruce Messinger, Madalyn Quinlan, Carolyn Houge, Stanley Rabinowitz

NOTE: This is a transcript of an oral discussion. Punctuation was inserted to help clarify.

Judy: The purpose of today's broadcast is to provide understanding and answers to educators who need to explain the CRT results to the public. Adequate Yearly Progress (AYP) results will be addressed at a later date. The format today is question and answer by a panel of six people.

The questions the panel is going to address today are from your e-mails to us. We have arranged and combined the questions. Should you have additional questions during the broadcast, we'll display the telephone number on the screen.

As you know, in the spring of 2004, you administered the first Montana criterion-referenced test. After the test was administered and you sent all of your student answer documents back to Measured Progress, several steps were followed; and I wanted to go over the steps with you so that you'll have that context. First there was preliminary item analysis and statistics by Measured Progress. Next (late June/early July), standard setting to determine the cut scores for each of the performance levels for each test in each grade were set by Montana educators. Those cut score recommendations were reviewed by our Technical Advisory Committee. (Our Technical Advisory Committee we refer to as the TAC. It is made up of psychometricians.) Finally those recommendations were sent to the Office of Public Instruction for the final decision on cut scores.

Other steps along the way were: the conversion of the raw scores to scaled scores, the results to schools and districts, today and the frequently asked questions which we are going to address, and the Technical Manual which will be out shortly.

First I'd like to introduce our panelists.
We will have three Montana educators asking questions:

Brian Hilton: Elementary and middle school principal from Ennis
Carol Wicker: Executive director of secondary education in Billings
Bruce Messinger: Superintendent of Helena Public Schools

Our panelists who will be answering the questions:

Madalyn Quinlan: Chief of Staff at the Office of Public Instruction

Carolyn Houge: Psychometrician and Assistant Director of Client Services at Measured Progress

Stanley Rabinowitz: Psychometrician at Measured Progress, and member of our Technical Advisory Committee

Frequently Asked Questions and Answers:

Brian: How do we know the CRT is aligned to Montana content standards?

Judy: In our Request for Proposals (RFP) and in our contract with Measured Progress, it was required that an outside and independent evaluator do an alignment study on the Measured Progress items and Montana standards. The Northwest Regional Educational Lab conducted the alignment study using the Norman Webb model. That study was forwarded to Measured Progress and then Measured Progress worked with that particular study to make certain that all of the standards were addressed.

Carolyn: The Montana Comprehensive Assessment System (MontCAS) also is constructed by Montana educators participating in national committees for item review, bias review; so they participate as well as others around the country in those discussions. And, as a matter of fact, we had a bias and item review committee meeting in September, and I believe there were at least seven Montana educators who went to Chicago for those meetings. So they're well represented at those meetings. So they have a lot of say about which items get put on the tests. Montana also reserves the right – another example of how they're aligned is that Montana reserves the right – to not score any items that appear on the test that they feel are perhaps out of grade range or something of that nature; so they reserve that right, should that ever come up. As Judy mentioned, there have been alignment studies looking at the content alignment. And I guess another example of how they're aligned with the Montana standards would be in the actual benchmarking, or pulling of exemplars, for the constructed response scores. Montana educators actually decide which papers from Montana will be used to score Montana students' results. So that's another example of how they're aligned. They're aligned with the expectations for student work.

Carol: Could you give us an explanation of what is standard setting, and why is it used to determine cut scores for proficiency levels?

Judy: Thank you, Carol. And I had asked Stanley and Carolyn to talk a little bit about standard setting. It seems to be something that we talk about. We talk about setting cut scores and standard setting and people use them interchangeably, which is probably fine, but we just want to make sure that everyone understands that. So Stanley, would you like to begin?

Stanley: The primary reason we give a state test is to determine whether or not the student has mastered the standards, has the student performed well enough. In classroom assessment, this gets done relatively informally. You may have a standard like 65 percent to 80 percent that you apply to most of your tests. Unfortunately, that method is too arbitrary technically or legally for something more formal such as a state test that has stakes for AYP and other kinds of instructional decisions. So we use a formal procedure called standard setting which is used to determine whether or not the test items – how many of the test items are needed for a student to match the standard. Now in a test like yours there are different levels – there's an advanced, there's a nearing proficient, there's

proficient – so we need to set that standard at each of those points. The bottom line is, can we look at the standards, look at the test, and determine through a technical defensible manner whether or not the student has met the expectations of the standards and Montana educators.

Judy: Thank you, Stanley. Carolyn, do you have anything to add?

Carolyn: I would support and echo everything Stanley said. Another way to think about it is that Montana educators, as I just talked about, had input into the content of the test; and now the setting of performance standards is a way for them to have input into the expectations about how good is good enough for Montana students on this test. So once you have aligned the content, you still need to make the decision about how good is good enough for certain levels. How good is good enough for proficiency; how good is good enough for advanced and nearing proficiency. So, those are all important Montana decisions.

Bruce: Judy, as the standard setting was described and how that's established, maybe you could elaborate further about the validity. In other words, when the standards set the level of difficulty to determine that the students demonstrate mastery on that standard. There have to be other reference points so it's not done in a vacuum or isolation. So maybe you could expand on that.

Stanley: There are basically four reference points that are used. The first starts with the content standards; and teachers, as part of the standard-setting process, need to understand what is the intent of the content standards. The second step is to look at the items themselves: Carolyn and Judy talked about the alignment study; this is the operationalization of the alignment study. So we look at the items and determine how well they align with those standards. That's the second step. The third step is to come up with descriptors. What are the different levels of performance, starting from proficient, moving up to advanced, and moving down to different reference points. In most states they have three or four levels of performance. The final step is to write descriptors. These descriptors say this is our expectation of proficient. It could be describing a proficient child. It could be describing what content standards are most important for that proficient score. Ultimately these four steps have to come together to lead to a valid standard.

Judy: Carolyn, do you have anything to add to that?

Carolyn: I'll just add one piece of information on to that. During the standard setting process, in addition to all of the steps that Stanley mentioned which are very content based and content driven and teacher experience driven, in the final phase of we also present them with what we call impact data. This is the percentage of students who would place in each performance level if they were to set their cut scores where they are proposing to set them. So it's a reality check in many ways. So there is this additional piece of information which is different from what Stanley is talking about. This is an external piece of information that is unrelated to the actual content.

Bruce: And would not the setting, establishing or determining the validity of items … Also, another methodology would be to do a comparative analysis on other like measures to determine if … You do the impact that you described, Carolyn, and then we have those same students take other like assessments. How would they perform on those? I didn't hear you describe that as a part of the methodology. Is that something that was done or could be done as a part of also establishing validity?

Stanley: What you're talking about is a step that typically happens following the standard setting process, the broader validation process. But it's a very tricky step. Let me explain why. The reason we give the Montana CRT is because it is different from other assessments. It's not the Iowa's; it's not a norm-referenced test (NRT); it's not a local test; it's not the SAT; it's not NAEP. Those are a lot of initials I'm throwing at you but there are a lot of tests out there. Ultimately you need to decide what is the right standard for the state, and how much these different tests should be correlated with one another. If they're all correlated perfectly, they're either measuring a very narrow thing or they're measuring the exact same thing; and that's unlikely. So when we look at the validity of a program like this, we have to determine how high is good enough for validation without being so high that it's redundant. Typically you want the CRT to correlate roughly between .6 and .8 with these external measures. What that means is: somewhere about half of the performance on one is similar to performance on the other, so that the correlations are all in the same direction. Meaning they're measuring similar constructs, but sufficiently differently so it justifies having this test. Ultimately, since the Montana CRT is the one most closely aligned with the Montana standards, that becomes the gold standard; the others become secondary checks for the program. It's important that you want – you want correlations, but you don't want too much redundancy.

Brian: Other than the Montana CRT, are there other tests that utilize standard setting?

Stanley: Whether you know it or not, all tests involve standard setting because there is a passing score and a non-passing score. What's different about the Montana CRT and other more formal standardized tests is the amount of effort that goes into the standard setting process: the convening of the committees, the training of the committees, the use of impact data, and the correlation with other measures. We certainly wouldn't expect a classroom teacher to do that on a Friday quiz, but we would expect the state to do that for something as high stakes as a state test. And therefore any test you spend money on ought to have a formal standard setting process; and the less you pay, the less you ought to depend on it. The more you pay, the higher the stakes, the more important the results, the more formal the standard setting process ought to be.

Carol: How is the standard setting process evaluated? How can you tell if it's a good process or not a good process, or sufficient process, or complete process?

Carolyn: What we do at Measured Progress is, in addition to observing and participating in the standard setting ourselves, so our staff runs the standard setting. For example, I was running one of the rooms for this standard setting; so in addition to that kind of feedback, we also have all of the participants fill out formal evaluations; and we talk to them

throughout the process. We ask them, "Do you understand?" There are constant checks for understanding as we go through. "Are you clear on what the task is?" But then in the end we ask them to fill out formal evaluations, and we do compile the results and report them. They will be reported in the technical manual. The results, I've seen them, were very favorable on the standard setting, as they typically are. Those results were presented to the technical advisory committee when we met in July as a piece of evidence supporting the recommendations coming out of that committee work. So those results were reported to them at that time. For the most part, the items are rated on a scale of one to five. Most of the participants rate threes, fours, and fives, and the average would be around a four. So they're very satisfied with the process.

Carol: So you're saying that is what our results are?

Carolyn: Yes, that is the evaluation on the Montana standard setting. Those results will be in the technical manual; and they are broken out by grade and content area, along with comments that were collected.

Stanley: Let me take this back a step. We'd like to make psychometrics look very difficult. It's good for us, but it really isn't necessarily that different from things in everyday life. How do you judge standard setting? How do you buy a car? The first thing you do is do research on cars. And that's what we do with standard setting. What methods work with tests like this? It's a good idea not to be first when a new edition of a car comes out. It's the same thing with standard setting. You don't want to be first. You want a method that's been proven to work; and the method that was used here has met that criteria. The second thing you do with a car is you test drive it. You see how it works. And the process that Carolyn described about how the standard setting takes place is similar to the test driving. You get in that car, you do it. Then you see how you feel driving it. And those evaluations Carolyn talked about are the same thing you do after you test drive that car; and the ratings were very positive here. The third thing that you do after you buy the car, you measure the results as you're driving it. You keep your gas mileage for the first six months; see if it's working well. You bring it in for tune-ups and oil changes over time, and that's the same thing that needs to be done as we go into the future of standard setting. Are the standards working? Do they have to be tinkered with a bit? Is it aligned properly? So I think there are metaphors in real life that can work even in psychometrics and testing, and I think that one might work pretty well.

Carolyn: That's a good analogy. With the standard setting process that we use, it really is all about the process itself. Judging how the process worked. Did the process flow? Do we trust the process? And if the process worked then the next question is, are the results reasonable? And that's when those results get presented to the technical advisory committee and to the department ultimately for reasonableness checks. So for the participants who actually – the Montana educators who actually participated in the standard setting – we focus very heavily on whether or not the process is working.

Brian: When the alignment took place and standard setting took place, was the level in which students would be at when the testing window would occur, was that taken into

consideration? For example, a student in fourth grade would basically be at 4.7 or 4.8 – with still a month and a half or two months left in school. Was this taken into consideration when alignment was made?

Carolyn: When the standards were actually set during the standard setting process, it is described to the participants as: what we need to do is determine what reasonable expectations are for Montana students at this point in the school year, when they took the tests. So yes, it is incorporated into the directions given to the panelists, the Montana educators while they're setting the standards.

Stanley: I'd like to follow up on that, too. There's certainly a desire to test as late in the school year as possible to make sure you get every minute of instruction into the process and that's why we test in the spring and not in the fall. But you also have to remember that these tests are cumulative. They measure cumulative knowledge over many years. And therefore even though it's the fourth grade standards, the fourth grade standards represent the culmination of fourth grade, third grade, second grade instruction and another week or two in that window. I mean we all think that if we had one more dress rehearsal for doing this or that we would be that much better. We may be a little better, but I don't think an extra week or an extra month is going to make up for three years of good or bad teaching before that. So yes, we do want to test as late as possible; but I think it's somewhat overestimated, the effect of an extra week or month of instruction. It might help on a couple of points, but it's not going to be the difference between a student who is way below proficient and above proficient. We want every bit of instruction – and I'm not saying that's not important, but I don't want that extra week of instruction or remediation or practice to interfere with the overall validity of the results which is a four year process, not even a one year or a one month process.

Bruce: One of the questions I wanted to ask about having to do with validity and reliability: Carolyn and Stanley, you both described a pretty tightly controlled process on a proven methodology of , and that's how we started the process. That's what I heard you describe, and now we're going to roll in thousands of more students. We're going to expand grade-level testing. So what is the strategy that will be used long-term to assure reliability and validity of this criterion-referenced test established by Measured Progress so that Montana educators have confidence that in fact, in their judgment, the results do reflect how their students are doing in their classroom? So what's the plan for that?

Stanley: We'll start generally and then I'll ask Carolyn to fill in more. There are usually three steps to a state validity plan. Reliability is relatively easy because ultimately reliability is a function of do you have enough items on the test and are they properly aligned to the blueprint. And from what I've seen as a member of the TAC, there are enough items and the alignment seems reasonable. So let me spend more time about validity, and let me talk about three steps.

First is what we call more internal validity, and that is the continued check that the test items are fully aligned to the standards and the blueprint. Every form of the test has to meet that standard. The fact that it happened two years ago doesn't mean that when you

add a fifth grade or sixth grade test that you've met that standard. You've got to do it individually for each test, and then you have to do it again for each form of the test if the test changes from year to year. Even if it's just a few items changing, you need to make sure that those new items are replacing the old ones properly. So that's the first check.

The second check comes down to the relationship of performance across years. It was hard when you had the gaps between grades because it's hard to predict what goes on from fourth grade to eighth grade and all the way through there. As you're building a fully articulated system from third grade through eighth grade and up through high school, it becomes easier then to determine that decisions are consistent from grade to grade, keeping in mind that students vary from year to year, teaching varies from year to year and, I'll tell you a little secret, even test results vary from year to year. We're not perfect at developing tests, that's why we have bands around test scores, so that's the second check, whether or not results make sense across years.

The third check is more global. Does the whole system work? And that's what we were talking before, do test results align with NAEP and SAT or ACT scores? Do the NRTs look similar to the CRTs. Not identical, but similar enough. If you have an instructional initiative – suppose you have a big third grade math initiative. Does that turn out in changing third grade or fourth grade math performance? So you start internal; you look at the tests themselves. You then look across the tests, and then finally you look across all different measures: teacher evaluations of students, other external measures, and just the sense (this becomes part science part art) that the test is working. It's showing progress; it's not showing unprogress.

And those three steps ought to be part of the state validation system.

Bruce: The predictive validity of the tests, is the methodology that would be used over time is – we'll look at last spring's fourth graders and look at that as a cohort group and we'll look and see how they do when they take the eighth grade test and see if there is a relationship there and I understand there's a lot of things that happen in four years; but part of what there's an interest in is whether the difficulty level of the test is consistent over grade levels, and whether is there a predictive validity with the assessment itself, so that when you look at thousands of students some of the differences will melt away statistically. Is that what I heard you say?

Stanley: I'm more comfortable with the fourth to fifth, fifth to sixth, fourth to eighth. It should make sense. It's more of an ordering effect than an exact match there, and that's what I would be looking at across years.

Carolyn: Actually, those students that took the fourth grade test last spring will take the sixth grade test in 2006, so there will be a smaller gap there. They won't actually have to wait until they take the eighth grade test to have an accurate measure on the MontCAS; and then at that point there actually will be yearly testing, so it will make it much easier to track longitudinally that proficient students are maintaining their proficiency, or I should say proficient students who are putting in and getting another year's worth of education

are maintaining their proficiency; and those students who were nearing proficiency and who've had some extra attention and are putting in extra effort are making proficiency – things you would expect to see.

Brian: I realize that there are a lot of questions out there regarding cut scores. How did the OPI make its decisions on cut scores, specifically regarding fourth grade math scores, and are there any plans to change those scores from 2004 in the upcoming 2005 scores?

Madalyn: After the standard setting process last June, and then the meeting of the technical advisory committee in July to look at the work of the Montana educators in the standard setting process, then those of us from the OPI who were part of the Technical Advisory Committee meeting brought a recommendation back to the state superintendent. Those folks included Judy Snow, Bud Williams our deputy superintendent, myself, and also Bob Runkel, from our special education division. What we presented to the state superintendent was a look at basically where the standard setting process had recommended that the cut scores be. So what's the break between novice and nearing proficient, nearing proficient and proficient, and then proficient and advanced? And each one of those recommendations from the standard setting process had a band around it which was a standard error. Well, basically the state superintendent chose cut scores at each grade level and in both subject areas that we believed minimized the risk that a student would be placed in a lower performance category than they belonged in. So we took a cut score that was at the bottom of the band around the standard setting process recommendation. So we took the lowest cut score that we could that was still within the standard error, in the belief that that would minimize the risk that a student was put in a lower performance level than they actually should have been put in. I hope that answers the question. There has been another question about whether Montana, whether the OPI, plans to change the fourth grade cut score. The answer is "no" for the current year. We know this is a work in progress; we know that we've got a lot to learn about all of this, and in a little bit we'll have some more discussions about what we might do in the future to try to validate or help us adjust the 2004 test results.

Carolyn: The only thing I would add is that as these discussions were taking place, Measured Progress was in communication with OPI very frequently – and at least some of the task members – to talk about the implications of placing the cut score in a certain position versus another position. So we did have lots of conversations about what are the implications of one route versus another. Because ultimately does come down to a policy decision, and as the OPI was taking the recommendation from the panelists and the recommendations from the technical advisory committee and synthesizing that into their policy decision, we worked with them to talk about the implications.

Carol: So then how will the cut scores be determined for grades three and five through seven in reading and math?

Carolyn: When those tests are brought on-line in the spring of 2006 – at that point, most likely – and this is our current plan, we will do a validation, a standard cut score validation type

of standard setting. So rather than starting from scratch, we would take the standards that exist in grades four, eight, and ten and we would use those as starting points. Typically that proceeds as establishing proposed cut scores for the intervening grades, and then bringing in panelists of teachers who then discuss whether or not those are reasonable standards in those grades. And potentially, at that point other grades could change. The goal there would be to smooth the standards across grades. So that you do have this coherent system where the proficient third grader who makes a year of progress is a proficient fourth grader, becomes a proficient fifth grader etc., so that the system is a line.

Stanley: Can we go back to just the broader issue on standard setting, deal a little more with the fourth grade math, and just how the other things are going to go on, bring some national perspective and just some reality checks to this process. We'd like this all to be science, but as Carolyn said this is not all science. There are technical pieces, there are human pieces, and there are policy pieces involved in standard setting. And it looks like, from my outside perspective, as though the state has gone through a reasonable set of balancing. Let me explain what I mean by that. As I said earlier in the conference, it starts with the standards and the panel's review of the standards. The way this process works is, different grade panels look at standards and items for that grade. They're the experts in their grades, so they make the best judgment they can make. And then Carolyn talked about impact data as the next step. And that's looking across the different grades and seeing if the standards make sense within and across grades.

What the TAC does is look at two things. My role on the technical advisory committee is to first evaluate that the process used was an appropriate process for the test and second that it was carried out properly. And from knowing what I know about the method that was used, and looking at the evaluations and hearing the descriptions by Measured Progress and the OPI, I'm reasonably certain that the process is correct and was carried out properly. But it doesn't stop there.

The next steps go through those technical and policy types of reviews; and it looks like, from what was described by Madalyn, the procedure used by the state superintendent is consistent with what other states do and practice. And here's what I mean by that. What she tried to do, as I understand it, is balance different kinds of information. The panel said "put it right here, and with a band around it." But the panel didn't say, "There's a band around it"; the panel said "We think this is the best guess." But because every method has error involved, this band gets produced just like a student score has a band around it, and so does the cut score. The TAC looked at two different types of error. They looked at the error of the standard setting process and the error of the test itself and said, "Given the two types of error, you could put it any number of places." So that in the standard setting process, the panel said, "Here's the error." The TAC said, "Well, there's other error too; you could actually put it here." But doing that would have too much overriding of the panel. So what the state superintendent apparently tried to do is balance what the panel said. They were the ones on the ground; you have to start with their judgment. The TAC said, "You could put it somewhere else," but the best information that was available is where the panel put it.

What Carolyn talked about in the future and what's going to be a lot easier to do in 2006 and 2007 when we have every grade filled out, is smoother for four to five, five to six, six to seven, seven to eight, all the way through. Nobody knows what the relationship should be for a test exactly. There are different tests and there are intervening activities that take place. So what the state superintendent did was for what I'll call a first-level smoothing. She took it as far as the panel said you should be able to take it. So she honored the panel, but still made an adjustment in the direction of smoothing, of making it more aligned, more coherent across grades.

What the state now has to do in my recommendation and in the TAC's recommendation is see if that was enough. Take the next couple of years. You don't know what the results are going to be in 2005. Maybe in 2004. … The reason you don't smooth immediately is this is a new program, a new test. You don't know if the fourth grade was actually right, and something went wrong. Teachers didn't quite teach the standards, kids didn't know the test, the alignment wasn't perfect yet. By just arbitrarily saying we're going to make everything the same in the first year, what you're guaranteeing is they won't be the same in the second year, the third year and the fourth year when this is a more mature program. So what I understand happened is that some smoothing took place in the direction of more coherence.

When this is a fully aligned three through eight/ten system and when people are more familiar with the testing, when students are more familiar with the standards, that's when you go for the full coherence, because right now the best information you have is the panel with the error adjustment moving in the direction of smoothing. And I don't know if you should wait until 2007 to make the next adjustment. If after 2005 and 2006 there still looks like a big anomaly between fourth grade math and all the other ones, I might just nudge you a little bit to make another adjustment after that. I don't know. If it's moving in the right direction, leave it alone. If it starts moving further away than what you want, or if the other grades start flip-flopping, then now you have two years of data. I trust two years of data a heck of a lot more than I trust one year. And I certainly trust three years in a mature program more than one year. So my recommendation is keep examining it. What you did was defensible, probably reasonable given it's the first year, and now keep monitoring as the program progresses. Anything more than that is just knee-jerking the other way, and now you have to fix it again next year. Which way do you do it?

Bruce: Stanley, I appreciate that explanation; and actually my question was whether there will be adjustments. I think you've answered that question in your advice. Your technical advice is to monitor that closely. I think the anxiety that educators are feeling across Montana, and I want to go back to the example you gave of a car. So maybe some models are released too soon. Maybe because we're forced to, or they feel the market demands we put a car out sooner than it's really ready. And in some ways we were forced to get this thing up and going. I mean, the reality is we didn't have ten years of history in Montana of a criterion-developed test just like you're describing – be great to have ten years of history on this – but I think the anxiety they feel is that we all know, you all know, that now they're held accountable for this. Even though what you just

described is a refining process that needs to occur with this test over time, and I concur, the anxiety is some people are now entering their second year of not meeting adequate yearly progress. The anxiety they feel is increasing rapidly, and so I appreciate your acknowledgment that this is a test that is developing and is maturing, and we'll know better in three or five years whether that fourth grade standard was the right level or not.

I think just acknowledging that's really important. And then together as educators we need to work through these rough waters, because there's some anxiety around that. Folks saying, "Is the test even right?"; and the reality is we'll know better about that over time when we have some longitudinal data to look back on. So I appreciate your comment, and I think just an awareness of that, an acknowledgment of that, and then we'll just work through this together, because – well there's no going back, and we have to have it. So I think the answer to my question would have been, will there be adjustments over time, and the answer is – it depends on what we find over time with the student's performance through the standard setting process and reviewing their performance – if I get it.

Stanley: Yes. So you're at school now, what should you do with these results? This is real important, and I want to differentiate things that happen to schools versus things that could happen to students. Fortunately this is a low stakes student test, so nobody is getting promoted (or at least from the state's perspective) from this test.

So suppose you have a fourth grade result for an individual student, and there's always error at the student level. I'd be careful about doing anything radical for that student based on one test, even the most perfect state test. You ought not to just change the entire program just because of that result. Let's just say the standard is off a point or two, I don't know. Maybe it's exactly right; I don't know. It could be too hard; it could be too easy. On the average, the panel said it's just right. What you ought to be doing at the student level is looking at fourth grade test results, looking at the student's performance across the entire school year and this test – and make a collective judgment from all of those data, particularly the teacher judgment of how that student had been doing all year, to decide what to do with that student in fifth grade and as we move on. So that's just good instructional local policy. Don't use the one test for anything that's going to affect the student too extremely one way or the other.

Now at the school level, there are also concerns that a lot of the data end up washing out. Sometimes it's too high, sometimes it's too low. Fourth-grade is just one grade. Eventually the system will have more checks built into it and the idea is to get through the rough stages now and understand them, explain them. The fourth grade cut score is the best data that you have now and you have to stick with it.

Next year you will have more data, and that's where the school judges, but don't over-evaluate, don't over-use the tests at the student level, because even if the test was perfectly aligned, perfectly developed with perfect cut scores, you still shouldn't do that at the student level, because it's a much more complex issue.

Brian: That makes sense, Stanley. We would like to utilize the tests as much as we could, so how can Montana educators compare the raw cut scores between fourth grade math and eighth grade and tenth grade?

Stanley: The reason that the test is produced in scaled score units as opposed to raw score is that an attempt has been made to make scaled scores relatively similar across grade levels. Now I didn't say identical. The reason I didn't say identical is that you don't have what we call in technical language a fully vertical scale. Meaning that with the Iowa's or other tests a 300 means the same thing whether it's second grade, third grade, fourth grade; scales are built consecutively grade by grade, so that exact meaning can take place. You have two forms of comparison that you can make from your current program. Not raw score but scale score broadly determined to be relatively equal; plus, I hope you use this instead, the performance levels. What I would be looking at now is the percentage of kids in each of the four levels that you have, and whether or not, within each grade, it makes sense to you to be doing that – does it makes sense that 40 percent of my students are proficient, and 20 percent are advanced? Does that make sense to me from what I know about my students, from what I know about other kinds of assessments I would be using? Intracomparisons right now because that's the best data you have is with your teacher tests, your teacher evaluations, the norm-referenced testing, other testing that you have that are more congruent, that are more consecutive to eighth grade or fourth grade scores. Fourth to eighth as we talked about earlier is a real big bounce – a lot goes on. I know my son when he was nine and when he was 13; he was a very different kid, educationally and otherwise. And therefore I'd be real careful interpreting right now fourth grade as a predictor of eighth grade. I would use it primarily as what I call a very loose early warning. Fourth grade is telling you what potential problems might show up in eighth grade. Eighth grade is a better estimate of what could be happening in tenth grade, but what I don't want you to think is fourth grade is destiny for eighth grade. Fourth grade is just ringing some bells and that's why we have fifth, sixth, seventh, and eighth grade. We expect some kind of intervention, instruction, remediation, regular teaching to take place. Use the fourth grade as a warning. Identify which students showed up in what category. Whether you've got gender differences; whether you've got some racial differences, however it plays out; whether you've got language differences. Interpret the data kind of in order. Do some groups score higher than others, and is that predictable? So use the data for what it can give you. Look broadly across the grade levels, but don't treat them as destiny.

Judy: We are having the results compiled and compared by subgroup and also by standard, by grade and by subject. So we can do many of the things you just mentioned statewide. The iAnalyze program that comes with the test allows districts and schools and classes also to look at those very carefully, and that would be a very positive use of the data.

Carolyn: Another advantage to having the results this year is that it allows you at the school level to look at your expectations that you have for your students and to gauge how they compare to the expectations that the state has set; so it gives you some idea of your expectations, a reality check on them in a sense. And perhaps not at your schools but at some schools this has been a wake-up call, where they thought their students were doing

very well, but they didn't have all this information about what the expectations were elsewhere in the state. So it can serve that purpose. It can also happen the other way, where you find that your schools did very well, and you feel that you are right on target and are doing very well, and you may be in a position to actually help some other schools understand better the kind of expectations you have in your schools. So there are all kinds of pieces of information that can be gleaned from this.

Bruce: Stan, I heard you talking about looking at intra-assessment comparisons. As many districts as there are, there are probably as many different kinds of assessments that go on, I mean, Iowa Test of Basic Skills is one, this is one, and there could be other criterion-based assessments that the districts are using. And Judy, I'm interested in your comment about conducting further analyses and looking at the sub populations in those comparative analyses. My understanding is that the state has access to only two data sets: one would be the Iowa Test of Basic Skills, and the other would be the Measured Progress CRT. I mean, is there a way to do outreach to other districts that might have other measures in place as a part; and I guess this gets down to a validity question, looking at trying to determine that judgment on that fourth grade that you talked about, Stanley. Maybe there are other measures in Kalispell or Billings or Townsend, maybe they have other measures that they have set expectations on, that they have confidence in, that do assess that and their expectations. It would seem over time that a broader collection of other measures. Because the state only has two is my understanding; one's the Iowa test and the other is this test. And the analysis of this over time, besides just looking at the performance on this test and doing that intracomparison … I mean, how do we get that data from the field? Because I know that schools are saying, **"**Well, this is what I know; this is the information I have about those fourth graders." But who cares besides us? How do we get that to someone who can do that intraanalysis? Is there a way we can do that as a state, because we don't have other state measures; we don't have other curriculum-based tests that we've used over time that the state has collected. Is there a solution in there somewhere that would help districts offer up that intra assessment that would be helpful in this analysis?

Judy: Certainly I've had lots of questions from people saying just like that – they have certain tests that they give in their districts and they have their results on those and they have their results on the Iowa test and now the results on the CRT and they are wanting to know how to pull all of those together to get the picture. I think the answer I usually give is that each one tends to measure something different so it's apples and oranges, but we do believe in multiple measures but then how to get the multiple measures to work together so that we get a complete picture that is cohesive and is very complex.

Stanley: Let me give some suggestions, including some suggestions of what not to do. I think it's important to know good things and less good things to do. And with everything, there is a certain amount of my own bias coming through, and smart people disagree with me on this; so I'll just put out my view and let others respond to it.

I don't think you need more end of year testing. I think between the CRT program, Iowa's the year they're given, other local tests that are given, I don't think that's what the

system needs more of. I think what the system can really benefit from is what we call more formative testing. More quarterly testing that's a little more formal than most districts do. Whether it's midterms, quarterterms or however you want to describe it. There ought to be a combination of public and private efforts; and I know that there are many private companies that would be happy to sell you test items, item banks, that would basically enable you to assess your students' progress towards the eventual year-end test. So there should be no surprises when that test is given using more formal measures than a classroom test, which may be very arbitrary and might not be linked to the standards. It could be the same test that was used 20 years ago and it's still coming off the mimeo machine. So what I recommend to states – and I just finished with a task force in another state – is a combination of private enterprises coming in (and the market is filling with them) and what I'll call the state clearinghouse, where the different districts with very little support can basically contribute items to a state clearinghouse linked to the standards and where some reviews take place. The best review would be that Montana teachers get together once or twice a year and do a quick and dirty alignment study - that this item seems aligned to the standard and is of reasonable quality – because you can have good items that aren't aligned, and you can have aligned items that aren't any good. So with a little bit of training, teachers can come together once or twice a year (or it can be done remotely, because these are not secure items) and basically come up with a clearinghouse of items that teachers have looked at that could be used to do these mini assessments, these quarterly assessments, these standard-based progress tests that can be administered locally using commercial software, or the state supplied software, or things you probably have in your district already. What you usually don't have, usually you have better software than items. So I'm describing a process that can get you better items that are as good as your software; and since it's not high stakes, the items don't have to be great; they just have to be good enough to serve this purpose. And I think with a little bit of working through the associations, OPI, or other kinds of mechanisms that are in place, you can get pretty good – all you need really is ten items per cluster, per standard – for an indicator. What if you had 20 to 30 – you can do pre- and posttesting; you can do progress; you can do mini tests. It would be relatively easy through existing structure and moderate review to come up with just the kind of system that I think you're looking for – that will help teachers, as opposed to administrators.

Carol: I would like to get us back to – I think – a gnawing question out there in the field. And I think that there is such a difference in our test results from fourth grade and eighth and tenth grades. And there are a couple of questions here I would like us to address. Does that difference mean that the fourth grade test is easier or harder, or does the percent of proficient students, the big difference between fourth and eighth and tenth grade, does that give us any indication about test difficulty, or what does that tell us – because of the very wide disparity between fourth grade and eighth and tenth grades. We all have a lot of fears that are built into those lower scores, and we'd all like to know exactly what does that mean. Was the test harder, was the score higher, what do you think happened there?

Madalyn: You're focusing in on math specifically?

Carol: Right. That's where everybody's big panic is right now, I think.

Carolyn: Right. And I think this is related to all of the things we've been talking about the last 25 minutes or so. The performance levels – when they were set by the panelists, by the Montana educators, the fourth grade group was not influenced by the work of the eighth grade group, and the eighth grade group was not influenced by the work of the tenth grade group.

Carol: So they were independent.

Carolyn: They were set independently; and that's the process, the bookmark process, and the way that that is used. So the results then get brought back to the technical advisory committee and that's when those kinds of discussions start to take place in terms of cross grade alignment and what Madalyn was describing earlier about there being error bands and decisions being made within the error band. For example, the fourth grade cut score could have been made lower and the eighth and tenth grade cut scores could have been made higher. So that would have produced a type of alignment. There's an impact there on the eighth and tenth grade data then, and we really haven't talked about that. And when you're talking about smoothing and consistency, there is a trade off. So what we have talked about is the fact that the fourth grade scores looked different than the eighth and tenth grade math scores; but if we were to bring those into alignment, your fourth grade scores would be higher but your eighth and tenth grade scores would be lower, so there are trade offs there. So to get back to your question about how we compare the fourth grade data to the eighth grade data, at this point I'm not sure that we really can. As we talked about, there are three grades in between that are not assessed, and we don't really have a feel for what is going on in those grades. So we don't know if the difference between fourth and eighth grades is due to true instructional and learning differences, and perhaps that's the case. It could be. I'm not sure that's real likely, but it could be. We don't know. But when we do test those intervening grades, and when we take a look at smoothing – at that point, when we look at three years of fourth grade math data, we will have a better idea of what's reasonable, and which is the correct direction to adjust if adjustments are even appropriate. I know of another state (where I live), Colorado, where in math it has been accepted that the fourth grade results in math … I mean, the pattern has been reversed. The fourth grade results are very high. Then slowly – the percentage of proficient students slowly dwindles, so that by the time you get to tenth grade there is a small percentage of proficient students. And that is the model that has been intentionally put forth. So there is a variety of things that could be decided once you have a little more data.

Carol: But making the assumption that the fourth grade test is too difficult is not appropriate.

Carolyn: No. As a matter of fact, if you look at – if you were just to look at the difficulty of the items on the fourth grade test, they were slightly easier for the students than they were for grades eight and ten in math.

Carol: And also making the assumption that the cut scores were too high and that's what caused the issue is not appropriate without more data. Is that what you're telling us?

Carolyn: I think that's true.

Stanley: Let me make three points that as an outsider I think it's the easiest for me to make in this group. That is, there's an assumption that the fourth grade test is too hard. I don't hear too many people saying, well fourth grade is right let's make eighth and tenth grade harder. It certainly doesn't work for AYP, but it's at this point that it could be just as likely that the fourth grade is the right result and eighth and tenth need to be adjusted. I'm not sure there's much stomach for making eighth and tenth harder. Fourth looks different so there's a sense fourth is wrong. Is it because it's two to one that eight and ten is higher; or is it because eight and ten make schools look better? It could be a little bit of both going on; but bottom line, we don't know. The second thing is, and this is a very interesting point, consistency is not necessarily right. If they all lined up, you would feel good about it. Look, our process works, everything lines up; but they could all be wrong. We get used to looking at things lined up and thinking, "Oh, that's good," but that might not be the reality. It could be – let me put a couple of scenarios out – it could be that instruction really is different in some grades. We've seen massive movements in some states to put better teachers in certain grades because that's where the state test was going to be, and therefore the grade before doesn't get as good instruction. But as we said earlier, these results are cumulative; so maybe some bias took place in instruction that we don't know. So it's very hard to really look at the one year and say that that consistency makes us feel better; maybe it shouldn't. Ultimately, more years of data – I'd feel better if after a second year and a third year the differences are still there and you can't explain them; then you make adjustments.

Bruce: I'd just like to make a comment because I've been in enough conversations since the data – since the test scores were released. The anxiety that is felt in part around the fourth grade goes back to an earlier comment that you made, Stanley, about the intra-comparative analysis, because the history in Montana in recent years on the Iowa Test of Basic Skills is just the opposite of what we see in this CRT. Again, it doesn't make it right or wrong, but the tension, the anxiety that folks are feeling is in fact what we could see using the rubric the state had established to demonstrate proficiency and we're looking at quartiles, we're looking at stanines, it wouldn't matter which way you look at the NRT in regard to student performance in mathematics. Clearly students did better on the Iowa Test of Basic Skills at fourth grade than they did at eighth grade, consistently by district across the state averages. I don't know that that's true by subgroup, but I would argue that it probably is. What we're seeing here is just the opposite of that. We're seeing our fourth graders not doing as well on the CRT and doing better as eighth graders. So it's just a flip of the intracomparison. So that creates anxiety; it makes it hard to explain. I think that part of what we've talked about today is this longitudinal work over time. Let's judge this as we get more data; that makes sense to me. I think what may be helpful in the interim is a more in-depth discussion that goes on, maybe facilitated by the office or associations, to say what could be contributing to that difference we're seeing between the norm-referenced test and the CRT. Let's be more specific than just saying they measure different things. That's a start; they do, but beyond that let's be more specific. What is it that the fourth grade CRT is measuring that is aligned with the standards, and is

appropriate, and is judged to be the right difficulty, but has little to do with the NRT, and may in fact be contributing to these differences. Now the NRT is measuring these things which were clearly not aligned to state standards, maybe inflating those scores unnecessarily. Maybe we begin to explain not how to adjust the scores and the levels, as much as trying to describe the real difference in those measures; so we have a better appreciation, and it's a confidence piece – that the educators in Montana have a confidence that this is a good measure, and that we can better explain to ourselves initially and then to the parents and students why we would see those differences. I see that as a next step we can consider doing so we can better explain those differences, because that's a huge source of anxiety in the state right now.

Carol: Especially when our SAT/ACT scores and so forth have consistently been high. It's almost saying to folks that you have to believe this one is true or that one is true, because they can't both be true; and that's kind of the questions we're getting.

Madalyn: I'm wondering if it isn't, when we looked at the fourth grade, the fact that a smaller percentage of students in Montana scored proficient or advanced in fourth grade math that there is at least an indication there that there is a greater gap between the expectations that Montana educators have for student performance and the actual performance levels. We seem to have a wider gap in our fourth grade math than we see in other places; and I'm wondering if you can speak to that.

Stanley: It could very well be, and you know panels do funny things, even when you ask them not to. Let's take accountability out of the equation for a minute, and I know that's real hard, and I promise I'll bring it back in a second. There are worse things you can do than over-identify in earlier grades. Because what you're at least making sure is that every student who is likely to be at risk later on is caught and gets the extra attention he or she needs. I'd rather have – remember we're taking accountability out of the equation for a second – I'd rather have overestimation of problems at the earlier grades and underestimation of problems at the high school. And maybe that's what – maybe the panel members put some of that bias into their process. You know what: I want to make sure that this test really does identify every problem of a child that could possibly come through the system, that's my obligation as a teacher. I don't know; I wasn't there. Your evaluations aren't showing that, but it's real hard; we told them not to, but its really hard to get in the minds of people that are thinking they are doing the right thing. I've seen standard setting in high school. Let me give two examples: I had one member that said everything in the standards is appropriate so the student should need 100 to pass; and in the same panel someone said well my kids they work real hard so I'm going to have a low standard because they work real hard. The method has a way of getting rid of the high and low scores, but you know people bring in their own expectations. You only have like

15 or so people there as part of that process. If you brought another 15 in, the results would move one or two points; that's why we put the band around it, so the panel only gets you so far and then, as my wife likes to say, the proof is in the pudding. You've got to actually give the test and see how well the panel did, how well the test did, and that's what we're talking about now. I don't know what the panel did. Measured Progress and OPI had a pretty good evaluation that most of the panelists did their job as was explained but everyone brings biases into the process.